Rich Transcription 2002 Evaluation

J. Fiscus, **J. Garofolo**, M. Przybocki, A. Martin, G. Sanders, D. Pallett, A. Le

May 7, 2002
Rich Transcription Workshop

NIST Vision

- Tightly-couple automatic speech recognition with higher-level language technologies
  - non-lexical information from the source signal
    - speaker ID, speaking rate, prosody, emotion, non-speech sounds, etc.
  - real-time integration of language processing technologies
    - Semantic, syntactic, contextual, world knowledge, and ASR
  - integration with video input when available
    - face recognition, lip movement, gestures, people movement, object manipulation
- ➢ Improved resources for human readability AND automatic content-processing technologies
  - Translation, Detection, Search, Extraction, Summarization, etc.

NIST National Institute of Standards and Technology

Rich Transcription
(Broadcast News Example)

Traditional ASR Output

```
tonight this thursday big
pressure on the clinton
administration to do something
about the latest killing in
yugoslavia airline passengers
and outrageous behavior at
thirty thousand feet what can
an airline do and now that el
nino is virtually gone there is
la nina to worry about from a.
b. c. news world headquarters
in new york this is world news
tonight with peter jennings
good evening
```

Enriched ASR Output

```
<speaker name="Peter Jennings">
<sent type=decl> tonight this
<proper_noun> thursday
</proper_noun> big pressure on
the <proper_noun>clinton
</proper_noun> administration to
do something about the latest
killing in
<proper_noun>yugoslavia
</proper_noun></sent> <sent
tu[e=decl>airline passengers and
outrageous behavior at <numex
val=30,000>thirty
thousand</numex>feet</sent> <sent
type=inter>what can an airline
do</sent> <sent type=decl>and now
that <proper_noun>el
nino</proper_noun> ...
```

Derived Human Readable Transcript

Peter Jennings: Tonight this Thursday, big pressure on the Clinton administration to do something about the latest killing in about the latest killing in Yugoslavia. Airline passengers and outrageous behavior at 30,000 feet. What can an airline do? And now that El Nino is virtually gone, there is La Nina to worry about.

Announcer: From ABC News World Headquarters in New York, this is World News Tonight with Peter Jennings.

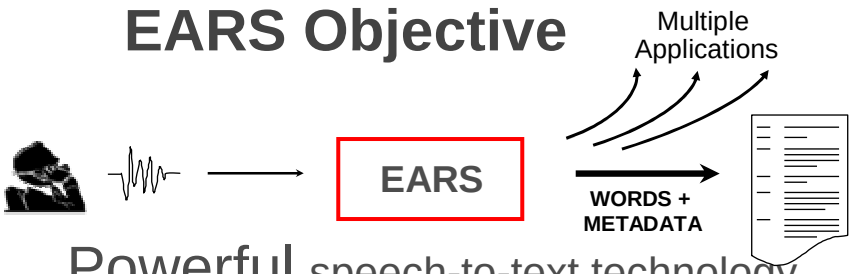
Peter Jennings: Good evening.

Annotated Word Stream → Human readable

Other language processing

NIST National Institute of Standards and Technology

EARS Objective



Powerful speech-to-text technology

Input: Human-human speech (broadcasts, conversations)

Output: Rich transcript (words + metadata) accurate enough for

Humans to read & understand easily

Machines to detect, extract, summarize, translate

Rich Transcription Evaluation Series will measure:

- Speech to Text Transcription (STT)
- Metadata Extraction (MDE)

NIST National Institute of Standards and Technology

RT-02 Goals

- Get community involved
  - begin studying STT/MDE integration issues
  - begin to determine MDE goals
- Baseline the state-of-the-art
- Begin building flexible/extensible evaluation paradigm
  - new formats
  - new software

NIST National Institute of Standards and Technology

RT-02 in a Nutshell

- Multiple Tasks:
  - Speech-to-Text Transcription and Metadata Extraction
- Multiple Domains:
  - news broadcasts, telephone conversations, meetings
- Multiple Channels:
  - two sides for telephone, multiple mics for meetings
- Processing Speed:
  - <= RT, <= 10X, >10X

NIST National Institute of Standards and Technology

RT-02 Evaluation Corpora

Broadcast News
- 60 minutes, 6 10-minute excerpts
  - systems could use whole shows for automatic adaptation

Telephone Conversations
- 300 minutes, 5 minute excerpts
  - systems could use whole conversations for automatic adaptation
- each conversation side in separate file
- from unreleased SWBD, SWBD II Phase 3, SWBD Cell Phase II

Meetings
- 80 minutes, 8 10-minute excerpts
  - systems could use whole meetings for automatic adaptation
- from 2 meetings per site: CMU, ICSI, LDC, NIST
- personal (head or lapel) mics/mix + omni mic

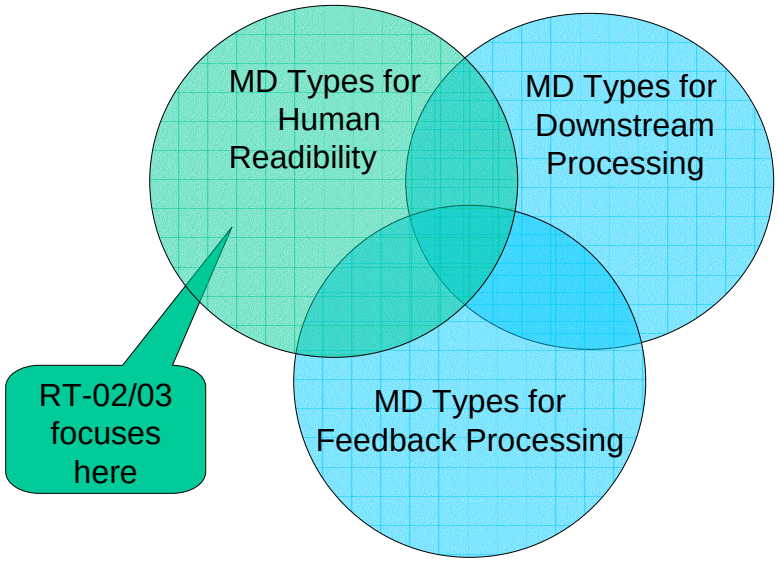
NIST National Institute of Standards and Technology

Speech-to-Text Evaluation

- Task:
  - Output orthographic transcript of speech in Hub-4/Hub-5 style
- Metric:
  - Word Error Rate
  - usual generation and normalization rules
- Due to SCLITE limitations, overlap was not evaluated, but **should be** in the future
  - broadcast news has some
  - telephone conversations have more
  - highly interactive meetings have over 50%!

NIST National Institute of Standards and Technology

EARS Metadata types?



NIST National Institute of Standards and Technology

What's the Right Metadata Set?

- Ran several experiments:
  - Suggested types from IBM, ICSI, LIMSI, MIT-LL:
    - Non-linguistically motivated:
      - speaker gender, simultaneous sound-producing sources, bandwidth, music, noise (vocal and non-vocal)
    - Linguistically-motivated:
      - Punctuation, capitalization, formatting, named entity, utterance boundary, disruption point, verbal edit, filled pause, quotation, parenthetical/aside
- Major Criteria for RT-02/RT-03:
  - Necessary information to derive human readable form Assumed dramatic script form and worked backwards
- Defer to later
  - Additional information to support downstream and feedback processing

NIST National Institute of Standards and Technology

RT-02/03 Metadata

- Obvious types of interest:
  - Speaker change detection/identification (RT-02)
    - permits association of speakers with orthography
  - Sentence boundary detection/classification
    - permits natural orthographic segmentation, capitalization, and punctuation
    - VERY difficult for spontaneous speech, will require research
  - Acronym detection and expansion
    - permits capitalization of acronyms and pointer to expansion
  - Verbal edit detection
    - permits removal of verbal edits for cleaned up transcript
  - Named entity/proper noun detection/classification
    - permits capitalization of proper nouns
  - Numeric expression detection/classification
    - permits numeric representation of numbers
  - Temporal expression detection/classification
    - permits natural representation of time/date expressions

NIST National Institute of Standards and Technology

Speaker Segmentation Evaluation

- Task:
  - Speaker Clustering: Find speaker change locations within an excerpt and assign within-excerpt speaker ID
- Metric:
  - Segmentation Error: measure of error of speaker assignments over time
  - doesn't measure accuracy of change locations

NIST National Institute of Standards and Technology

RT-03

- Same domains, multi-lingual
- New test set selection/scoring paradigm for inter-test comparison
  - larger test sets
  - filtered test subsets, mothballed systems, static and variable test epochs
- Increased metadata types
  - focus for RT-03 on those necessary for human readable transcripts
  - focus on additional types in later years
- All speech scored, including overlap
- New evaluation software
  - generic for major task types and extensible for new tasks
- MORE TIME

NIST
National Institute of Standards and Technology